Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Creating Wallpapers using Mobile Phone photos

A study on selecting and cropping photos automatically

Semester Thesis

Seonwook Park

Computational Science & Engineering MSc
Department of Mathematics

Supervisors: Michael Gygli and Dengxin Dai
Professor: Prof. Dr. Luc van Gool
Computer Vision Laboratory
Department of Information Technology and Electrical Engineering

July 28, 2015

Abstract

Mobile devices with cameras now contain various photos ranging from natural scenery and city skylines, to street signs and restaurant menus. When deciding to review these various moments in daily life, one can opt to use a wallpaper application which sets an image from the photo collection as a wallpaper. Unfortunately, not all photos are suitable nor well composed. This study is a novel attempt to improve these aspects. This is done by first deciding if an image is suitable to be used as a wallpaper, and if so how it should be cropped and shown on a given display. Qualitative results are quite good where photos with less desirable objects are omitted, and more interesting regions are retained in the final crops. The selection algorithm yields a classification error as low as 3.7%, and the cropping algorithm yields an improved 0.782 median maximum overlap score. While the selection part has no prior work, the cropping algorithm is an improvement over previous implementations.

Contents

List of Figures

Chapter 1

Introduction

The widespread use of mobile computing devices such as smartphones and tablet computers has led to the creation of large personal photo collections. Unlike previous methods, photo capture using modern smartphones has a low cost. There are low space restrictions and photos are simple to take with modern applications. This leads to lower inhibition in taking photos, and a wider variety photos in less carefully curated collections. These collections can include photos capturing among other objects:

- Natural scenes
- Cityscapes
- Quick notes (street signs, maps, documents)
- Screenshots
- Quick shots for instant messaging

Not all photos taken are of high value to a user. For example, photos of posters, street signs or maps are used for short term memory and do not have value in being reviewed. Additionally, some photos are intentionally taken with less care for use in instant messaging.

Photos which are taken with care can often have high value in being reviewed occasionally. To this purpose, applications such as Muzei for Android allow the displaying of photos from a mobile phone album as a wallpaper. Unfortunately, this is not done in the most ideal manner. There are two big issues, (1) not all images are appropriate to be used as a wallpaper, and (2) the image is not well aligned when displayed. For example, it is undesirable to have a passport scan as a wallpaper or an image where only the rear of an elephant is visible.

This project aims to improve this experience and other similar experiences by first selecting photos which are appropriate to be used, then cropping the photo appropriately to display on the screen of a mobile phone. This is to be done in a fully automatic manner where no user intervention is required.

Chapter 2

Related Work

As mentioned in the previous section, the aim is to improve the experience of viewing photos from a mobile phone gallery as wallpapers. This requires the determination of whether an image should be selected for display, and the selection of a window or crop of the original image which should be displayed on a given screen.

There is no prior work for automatically selecting wallpapers. The most similar work is the summarisation of photo albums [4]. Unfortunately, there is lack of detail and reliance on detailed annotations. Thus in this study, the problem of deciding if a photo could be used as a wallpaper is dealt with in a simple manner which requires minimal annotations.

Depending on which object or scene is given focus, an image can be perceived very differently by a viewer. Thefore there is great interest in attempting to improve how well an image is composed. To this effect, there have been many previous studies in cropping or deforming images to a target size or aspect ratio.

Seam carving is an effective way to removing less interesting seams or regions in an image but unfortunately can warp the image badly when failing to work successfully [5]. Other similar methods such as Multi-operator retargeting also suffer from heavy deformation and artifacts depending on the input image [6]. In fact, [7] notes that: *"Cropping, although a relatively naive operation, is still one of the most favored methods, most often since it does not create any artifacts. Our findings indicate that the search for an optimal cropping window, which was somewhat abandoned by researchers in the past few years, could often be favorable and should not be overlooked."* Therefore, cropping algorithms are focused on for choosing how to display a candidate wallpaper image on a given display.

The automatic cropping of an image can be done in two major ways, one which requires a set of rules, and a learning-based model which associates a set of features to a score. Rules-based croppers such as [8, 9] can suffer from bias or imprecision due to human-selected rules and parameters. Learning-based croppers such as [2, 3] are not without fault however, with low precision due to lack of training data being a particularly big issue.

A novel method suggested by Fang et al. [1] attempts to combine the advantages of both approaches by introducing three cues into the learning and cropping stages. These include saliency composition, boundary simplicity, and content preservation. This is described in greater detail in the next sections. Another improvement in the suggested method is the use of public datasets such as those which can be found on image hosting services such as Flickr and Photo.net. These services can indicate how good an image is perceived to be and allow for automatic annotation of possible image crops.

Chapter 3

Materials and Methods

This project consists of two parts. When given a set of photos taken using a mobile phone, the first part is the selection of photos and the second part is the cropping of photos. The selection criteria is wallpaper suitability while the cropping aims to retain interesting areas to result in a well composed final image.

Both decisions are subjective by nature. If a set of rules were defined to make the decisions, the process would inherently be biased, and results difficult to evaluate. Thus for both selection and cropping, machine learning is utilised with aims to avoid introducing unnecessary bias.

3.1 Selection

When considering a single photo, whether this photo could be used as a wallpaper is a subjective decision. One person may prefer city skylines while another may prefer distant mountain ranges. A simple way of assessing a photo in this case is via object class detection.

One way of performing object class recognition is via the use of deep convolutional neural networks. An existing implementation is one provided by Caffe, a deep learning framework. Caffe provides several models trained for the ILSVRC 2012 dataset [10]. This dataset consists of 150,000 photographs and 1,000 associated object classes.

In a deep convolutional neural network, different convolutions are performed at each neuron with a focus on different portions of a given image. In [11], it is determined that using the activations of the hidden fully connected layer, `fc7` in conjunction with a SVM results yields the best classification results.

Therefore, an input image is propagated through the trained neural network and the activations at hidden layer `fc7` is used as the given images representative feature vector. The exact neural network used is `bvlc_reference_caffenet`. This results in 4096 features per image. These feature vectors are used to train a SVM. The decision function of the trained model calculates a suitability score when provided a feature vector representative of an input image. If this score is positive, an image is suitable.

The resulting model can then be used to classify new images. Figure 3.1 shows the full pipeline for checking the suitability of an image as a wallpaper on a mobile device.

This portion of the project is implemented using `Python`, `scikit-learn`, `OpenCV`, and `PIL`.

3.1.1 Datasets

Two datasets were created for the selection step. I will call these datasets the *M*ichael dataset and the *W*ookie dataset.

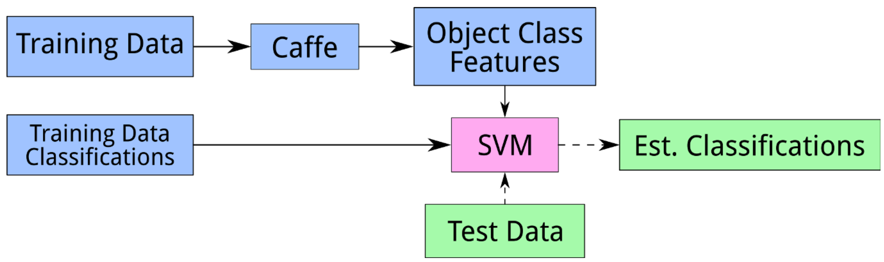


Figure 3.1: Full pipeline for determining if an image is suitable for use as a wallpaper.

The Michael and Wookie datasets consist of 275 and 266 images each respectively. The photos image a variety of objects and scenes with some example objects as listed below:

- Natural landscapes (Mountains, rivers)
- Man-made landscapes (Cityscapes, landmarks)
- Text (Poster, presentation, street sign)
- Dim indoors (Concert, restaurant, presentation)

While the photo collections were created with aims to provide sufficient variety to result in a more general purpose model, privacy was respected and thus the datasets lack people.

The ground truth was collected by annotation from 4 different volunteers. Each person was asked to mark images in both datasets as either suitable or unsuitable (1 or 0 respectively) based on the simple rule: *"rate as suitable if you would personally use the image or a portion of the image as a wallpaper on your phone"*.

In early annotations, it was evident that the order in which images was displayed affected final decisions. If similar images were shown in succession, the classifications could become less consistent. The order was therefore randomised. A simple Python script was written to accommodate this effort. Key bindings were added to allow for effortless annotation and navigation between images, and the images are tinted green or red depending on the decision to allow for quick review.

A further improvement could be made by splitting annotators into two groups, one which make a preliminary set of annotations which can be used to create a balanced dataset which is annotated by the other group of annotators. This would reduce bias further.

3.2 Cropping

To crop an image automatically, it should be possible to find out which regions of the image are worth retaining. To do so, we consider visual saliency. Visual saliency is a measure which represents how distinct a pixel is in relation to its neighbouring pixels. Many saliency map algorithms have been suggested in the past, with ground truth collected by tracking the eye movements of human participants.

The algorithm selected for generating visual saliency maps is the boolean map based approach or BMS [12]. This approach randomly thresholds each channel in CIE Lab colour space to generate a set of boolean maps, then averages the boolean maps to generate an attention map. The algorithm is very fast and produces high quality saliency maps. The output saliency map is further dilated and blurred in an attempt to give crop candidates a good margin from objects.

An input image is first scaled to be at most 800 pixels wide or high, then a dilation width of 2 pixels and step size of 6 is used to generate an initial saliency map. This map is further processed with a dilation filter of width 5 pixels and a 11×11 gaussian blur filter with $\sigma = 50$.

In [1] automatic cropping is done using three distinct metrics. These are saliency composition, boundary simplicity, and content preservation. Saliency composition concerns the layout of saliency energy, and boundary simplicity is whether the crop cuts through objects, while content preservation is the proportion of saliency energy kept in the crop. In our approach saliency composition and boundary simplicity are encoded into the learning stage where an SVM is trained to distinguish between well and badly composed images or crops. The content preservation metric is used in the filtering of crop candidates in the cropping stage.

Saliency composition is represented by a 4-level spatial pyramid of the saliency map. This is done by resizing the map into 8×8, 4×4, 2×2, and 1×1 patches by averaging pixel values, then using pixel values in these patches as features. This results in 85 features which encode the distribution of saliency energy.

Boundary simplicity aims to encourage crops which do not cut through objects. This can be done by taking a gradient map of the original image. A 2-pixel wide strip is taken for each edge and the mean value is used as a feature. This results in 4 features encoding the amount of edges crossed by a given crop's boundary. This is because left or right edges may require higher boundary simplicity than top or bottom edges.

The gradient map of an image is created by first resizing the image to be at most 600 pixels wide or high, then applying a first-derivative 5×5 Sobel filter. Absolute values are taken per pixel, then a 11×11 Gaussian blur filter ($\sigma = 50$) is applied. This results in an image similar to the corresponding saliency map where object boundaries are blurred to encourage good margins in crop candidates.

When considering saliency composition and boundary simplicity, the final number of features used to represent an image is 89. These features are used to train a SVM. The method of annotating crops as well or badly composed is outlined in section 3.2.1. The SVM trained is C-SVC with a linear kernel where 20-fold cross-validation is used to find the hyperparameter C.

The trained model can be used to assign a score to a candidate crop. For any given image, thousands of crop candidates are generated and evaluated to find the best crop. Specifically, 4000 initial crop candidates are generated, and a content preservation score ($S_{content}$) is evaluated. The score as given by equation 3.1 represents how much interesting information is retained by the suggested crop. By using this score, one can discard inappropriate crops early on without comparing scores given by the trained model. Crops below a threshold score are retained in a list of crop candidates. The threshold is reduced until a target number of crops is reached. This algorithm is described in algorithm 1.

The shrinking threshold encourages larger crop windows. This is quite intuitive when considering human attention which considers the whole image and focuses into smaller details to find a better defined area of interest.

The final list of candidate crops are then used to calculate a score representing how good the crop is. This is done using the previously trained model. The crop with the highest score is considered to be the best crop and is finally used to create a final crop of the input image.

This portion of the project is implemented using `C++`, `Python`, and `OpenCV`.

3.2.1 Dataset

A strength of the algorithm suggested by Fang is that the dataset used to train the SVM does not require human annotations [1]. This was accomplished in the original study by taking top images from Photo.net as images of class 1 (well composed), and taking random crops of these images as class 0 (badly composed).

Algorithm 1 Caption

$saliency \leftarrow$ Saliency Map
$thresh \leftarrow 0.7$
$n \leftarrow 0$
repeat
    Generate 4000 crop candidates.
    for all crop **do**
        $cropped \leftarrow saliency(crop)$
        $S\_content = sum(cropped)/sum(saliency)$
        if $S\_content < thresh$ **then**
            Add crop to candidates
            $n = n + 1$
    $thresh = thresh * 0.98$
until $n > 80$

A very similar approach is used in this study. 2000 top images from several subreddits of Reddit are acquired. The used subreddits are: CityPorn, EarthPorn, itookapicture, photocritique, WaterPorn and windowshots[1]. An advantage of using these sources is that there is great variety in the images, and the quality is quite good due to crowd-sourced selection. However, there is a bias towards natural landscapes as EarthPorn is the most popular subreddit.

The badly composed images are created by randomly generating crops of a well composed image. When this is done in a completely random fashion, the accuracy of the final algorithm varies greatly. Therefore two parameters are introduced, $T_{content}$ and T_{area}. $T_{content}$ is the minimum ratio of saliency energy within a proposed bad crop, and T_{area} is the maximum area ratio between the crop and original image. This is shown in equations 3.1 and 3.2 where S_{crop} is the saliency map of the crop candidate, and A_{crop} is the area of the crop.

$$S_{content} = \frac{S_{crop}}{S_{full}} \tag{3.1}$$

$$S_{area} = \frac{A_{crop}}{A_{full}} \tag{3.2}$$

$T_{content}$ and T_{area} are set to 0.2 empirically.

3.3 Full Pipeline

An example final pipeline combines the selection and cropping parts. When provided a collection of images sourced from a mobile phone, the pipeline should first select images which can be used as a wallpaper. When this subset is found, some diversification can be attempted such that a user does not view similar images in succession. This can be done using Maximal Marginal Relevance (MMR) which uses uniqueness and novelty measures to find the next item [13]. In our case, uniqueness can be calculated as the distance

[1] `https://www.reddit.com/r/CityPorn+EarthPorn+itookapicture+photocritique+WaterPorn+windowshots/top/?sort=top&t=year`

between the feature vector of the current image and a candidate image (equation 3.3). Novelty (N_j) can be represented as time elapsed since an image was last shown.

$$D(I_i, I_j) = \|F_i - F_j\|_2 \tag{3.3}$$

$$MR(I_i, I_j) = \lambda D(I_i, I_j) + (1 - \lambda)N_j \tag{3.4}$$

The index of the next image to show is then $\arg\max_{j \in \mathcal{J}} MR(I_i, I_j)$ where the current image index is i, all candidate image indices are in set $\mathcal{J}$, and $\lambda = 0.3$.

Chapter 4

Qualitative Analysis

4.1 Selection

Figure 4.2 shows 21 sample images from the Michael dataset. These images, as mentioned before, can include undesirable objects such as text, street signs, or store items. As mentioned in section 3.2, a SVM is trained using features representing object classes. When classifying the given images using this SVM, the resulting selection of wallpaper candidates are shown in figure 4.3.

It can be seen qualitatively that images with undesirable objects have been discarded, such as the image of furniture, a poster, or the model number of electronic equipment. Though in general photos with obviously undesirable objects are discarded as appropriate, photos of city skylines or prominent single foreground objects for example are not usually classified as being appropriate. This is due to disagreements between annotators. A larger number of annotations per image may help in this case.

When evaluating scores for all images in the Michael dataset, it can be seen in figure 4.1 that images of distant natural landscapes attain high scores. Unlike scenes such as dark indoors and cityscapes, natural landscapes tend to be classified as suitable by more annotators.

It should be noted that a single user's preference could be greatly different from most other users where the purpose and intent of having a photo wallpaper may differ. For example, user A may desire to have dark and slightly artistic photos mainly biasing towards indoor club scenes and long exposure shots at night. User B might be a big fan of album art or concert photos, desiring objects or scenes that we tend to assume to be undesirable. It was actually the case in this study that using greatly conflicting annotations caused issues in classification, where an annotator heavily favoured dark scenes with low detail or discernible objects.

Further work could be done to perform a weighted average of multiple classifiers depending on wallpaper preferences.

Figure 4.1: Top 7 suitable images in descending score order (Michael dataset)

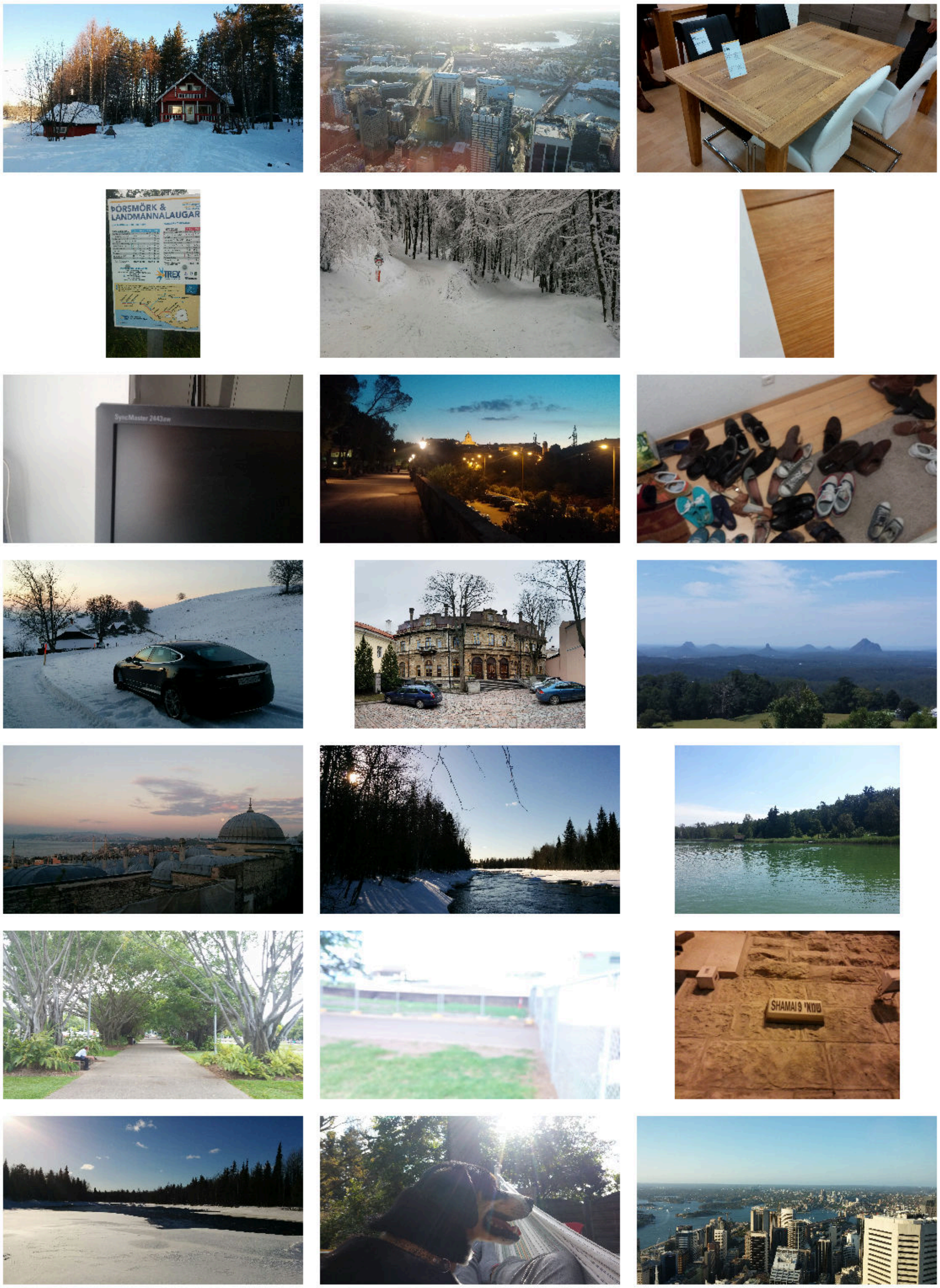


Figure 4.2: 21 sample images from Michael dataset.

Figure 4.3: 21 sample images from Michael dataset with images not selected dimmed and in grayscale.

4.2 Cropping

The automatic cropping algorithm works quite well in general. In the following figures, four images are shown per input image: the original image, a saliency map, a gradient map, and the final cropped image. It should be noted that brigher regions in a saliency map represent more visually distinct regions, and brighter regions in a gradient map represent regions with large changes in colour.

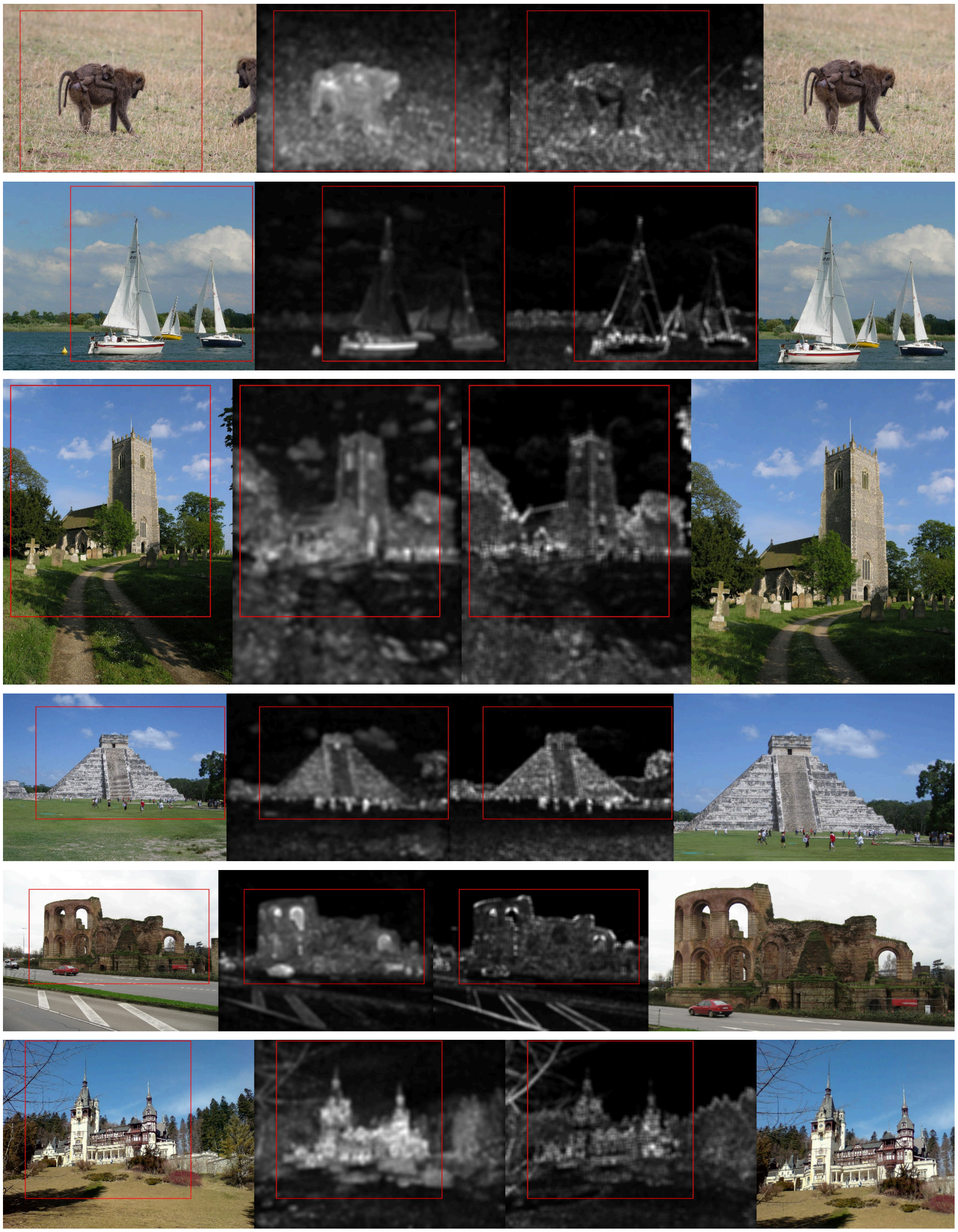
It is hoped that the saliency map emphasizes prominent objects well enough for irrelevant or less important details to be left out of the final crop, and the gradient map exhibit low values for background regions. For the cases where this is true, it can be seen that the algorithm works very well.

In figure 4.4 in particular, it can be seen that the most prominent object is retargeted well. This tends to result in better composed final images. Some of the crops even exhibit some considerations of boundary simplicity. Figure 4.5 shows this effect better. The classifier favours having lower gradient values on crop borders, leading to crops which do not tend to intersect objects. It should be noted that since the input gradient map is intentionally not normalised, weaker boundaries can be included in crop boundaries. As opposed to [1], the weighting of boundary simplicity is not done using a fixed variable but by relying on the training of the SVM.

The automatic cropper does have its pitfalls however. The most common error occurs when the final crop cuts through distinct foreground objects. This is shown well in figure 4.6. It can be seen especially well for the case of the first image that the fault may lie in the saliency map implementation used. Other background colours and patterns are deemed more salient at times making it more challenging for the algorithm to retain relevant but ”not salient” regions.

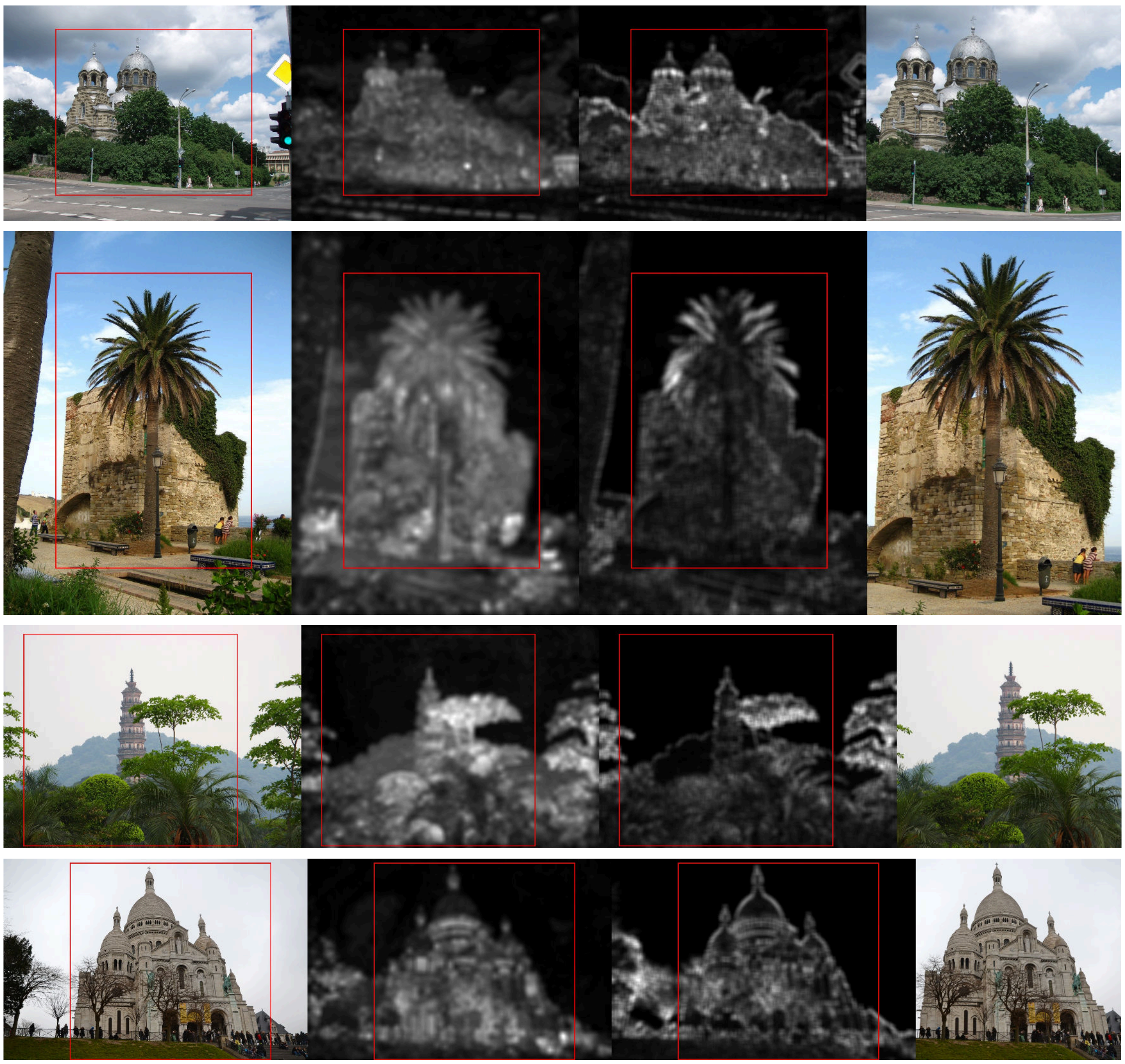
Another example where the saliency map implementation may cause issues is when reflections occur in an input image. Figure 4.7 shows an example with a flock of flamingoes and their reflection. Both the flock and their reflection is considered salient and thus the final crop is not composed as well as the initial image. This could be a cause for confusion for human croppers as well however, especially when a reflection could be considered artistic and thus should be well focused.

Overall, the automatic cropping algorithm works very well. Any mistakes could be improved in the future with better saliency map implementations.



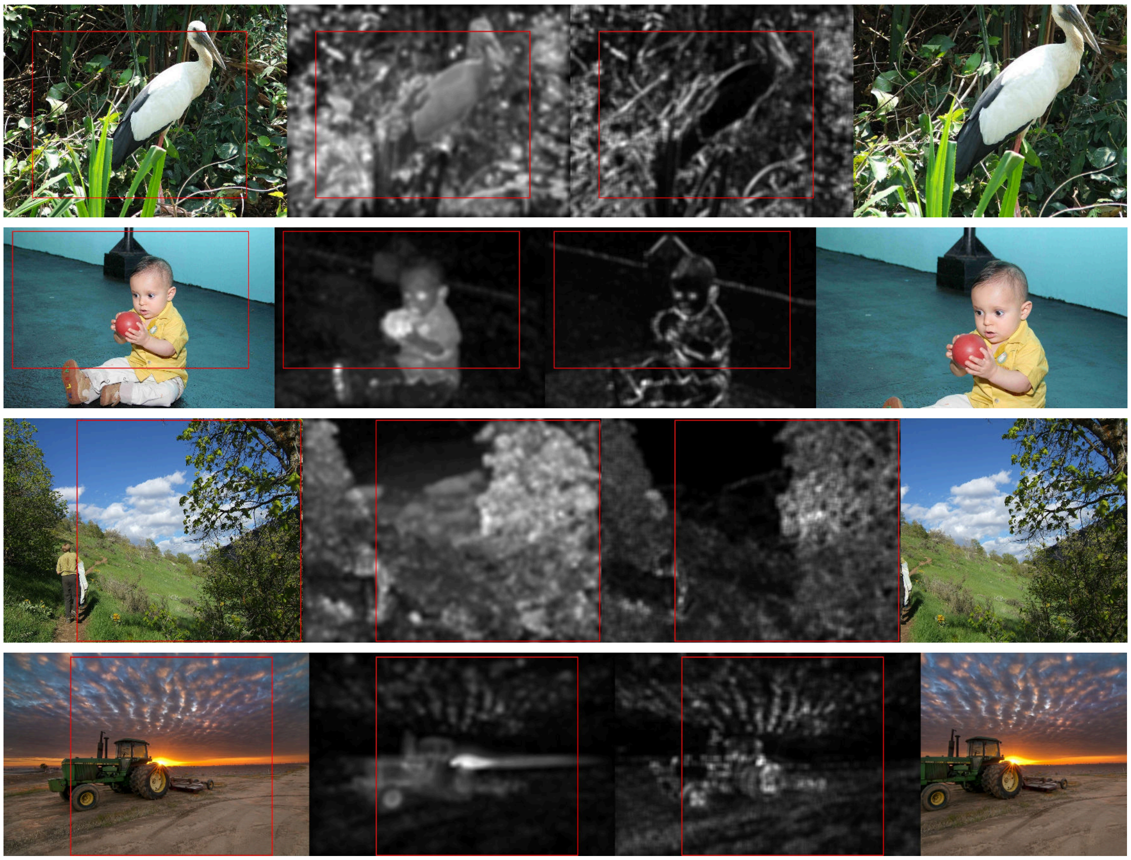
(From left to right: original image, saliency map, gradient map, final cropped image)

Figure 4.4: Crops with main objects isolated and centered.



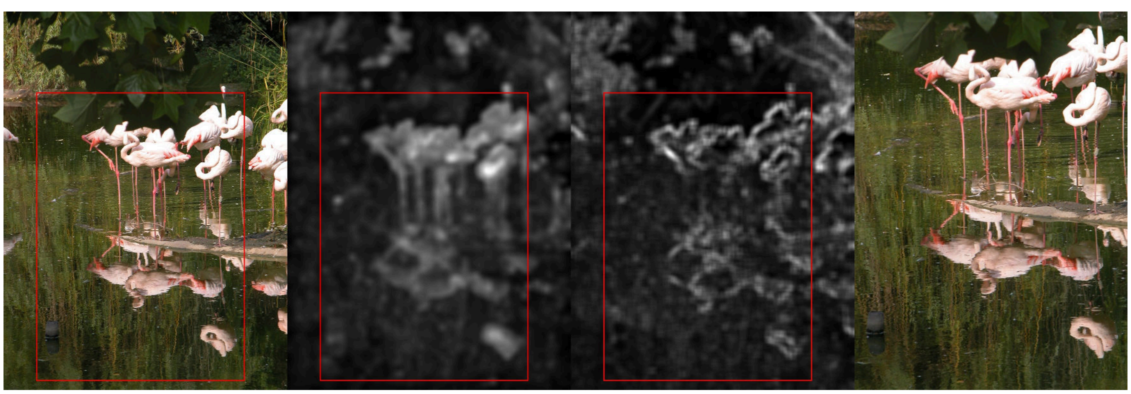
(From left to right: original image, saliency map, gradient map, final cropped image)

Figure 4.5: Crops with good boundary simplicity.



(From left to right: original image, saliency map, gradient map, final cropped image)

Figure 4.6: Unideal crops which cut through objects.



(From left to right: original image, saliency map, gradient map, final cropped image)

Figure 4.7: Unideal crop due to reflections.

Chapter 5

Quantitative Analysis

5.1 Selection

4 volunteers were asked to annotate the wallpaper datasets. For the Michael dataset, there is a correlation coefficient (R) of 0.49 in annotations, while $R = 0.28$ for the Wookie dataset. It could already be seen that the Michael dataset is perhaps of higher quality in terms of variety of objects imaged and lower repetitions.

When training a model on the Michael dataset and evaluating the model on the Wookie dataset, there are 5 incorrect predictions for 135 images where annotations agree. This is an error rate of 3.7%, much lower than the opposite where training is performed on the Wookie dataset and evaluation done on the Michael dataset. In this case, there are 42 incorrect predictions for 141 images where annotations agree. This is a 29.8% error, confirming that the quality of the Wookie dataset could be improved.

Furthermore, we evaluate the performance of the classifier when training on ones own annotations vs training on all available annotations. While this error fluctuates for each annotator, the average error is 59.2% for the case where training is performed using all available annotations. This is a much higher error compared the case when training is only done using an annotators own annotations (29.2% error). An interesting study would be to see if combining annotations with high correlation helps to improve the performance of the personalised classifier of a user.

Another analysis done is the assessment of the Precision-Recall curve of the classifier (figure 5.1). This is for the case when training on all annotations for either datasets and evaluating on the other dataset. For the case where training occurs on the Michael dataset, there is higher precision for low recall (< 0.4) cases. This may be useful for a conservative system where it is more preferable for the classifier to be correct than to make use of as many photos as possible. In general, both exhibit high precision.

5.2 Cropping

It has been shown that the automatic cropping algorithm works quite well in practise. However, it is difficult to tell if this is an improvement over the implementation by Fang [1]. The differences between the proposed algorithm and that of Fang are:

1. 4 separate boundary features used in training as opposed to a single boundary simplicity score post-classification.

2. A shrinking crop size algorithm for candidate crop generation.

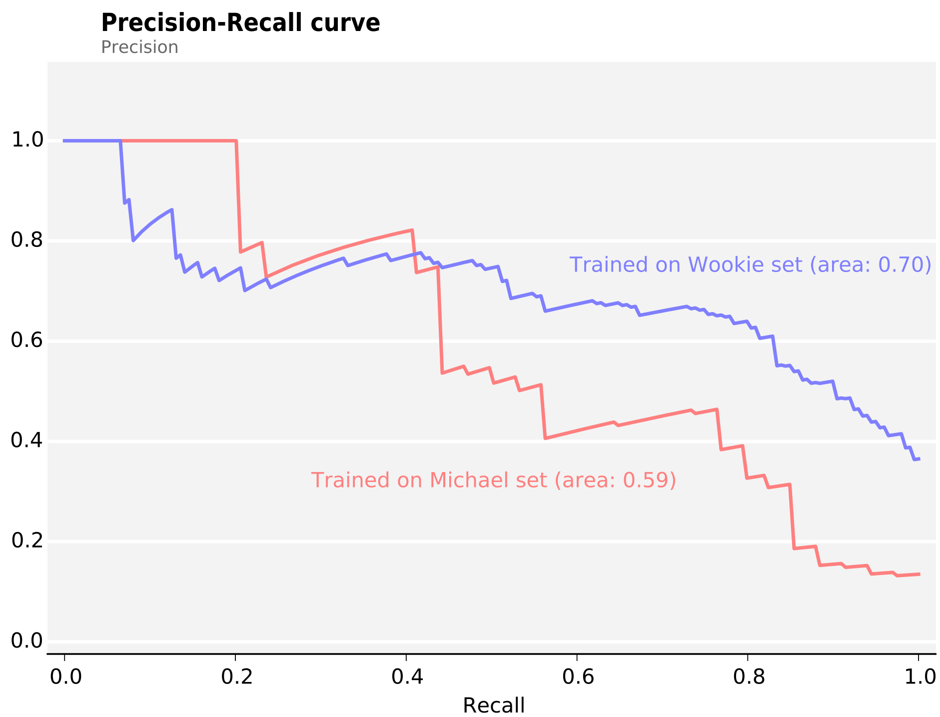


Figure 5.1: Precision-Recall curve of selection classifier trained on either dataset.

3. The use of a Reddit-based dataset for training.

To evaluate the effect of these changes quantitatively, the evaluation method introduced by Fang is used. Fang provides a 500-image dataset annotated with the use of Amazon Mechanical Turk. Each image has 10 associated ideal crop which can be regarded as a ground truth crop. This human crop dataset is used for quantitative evaluation.

Given a candidate crop C_i, the maximum overlap between the crop and available human crops is calculated as shown in equations 5.1 and 5.2. This can be assessed for just the top crop candidate as well as up to top 5 crop candidates. It is expected that the maximum overlap increases with more top candidates considered as the model is not perfect and the true top crop candidate may be a few places offset.

$$\text{Overlap}(C_i, H_j) = \frac{C_i \cap H_j}{C_i \cup H_j} \tag{5.1}$$

$$\text{MaxOverlap}(C_i, H) = \max_j \text{Overlap}(C_i, H_j) \tag{5.2}$$

The maximum overlap scores over top 5 crop candidates is calculated over the mentioned 500-image dataset to yield scores as seen in figure 5.2. It can be seen that the suggested algorithm works better in general. The standard error of maximum overlap values are negligible and therefore it can be seen that the proposed implementation is an improvement over [1]. Consequently, compared to [2] and [3] the top 1 candidate score is a marked improvement. Compared to [3], our method is almost a 2x improvement. It should also be noted that the maximum overlap score increases slower for our method than compared to earlier methods. This indicates that the top crop candidates are more reliable.

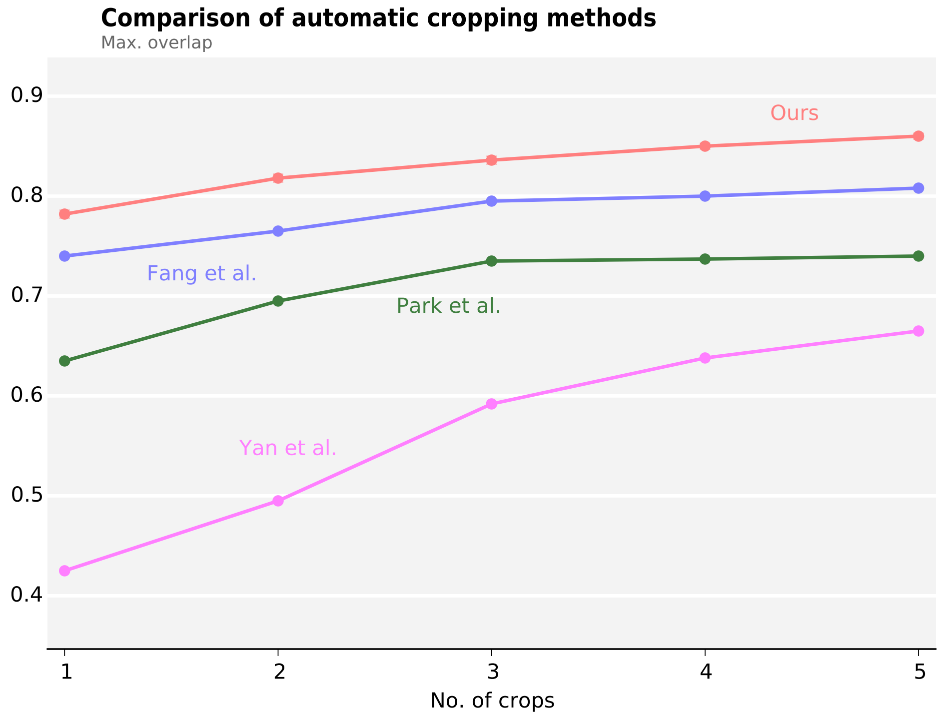


Figure 5.2: Quantitative evaluation of various automatic cropping algorithms [1, 2, 3].

In terms of MaxOverlap values, our implementation yields 0.782 ± 0.004 when top 1 crops are considered with the value increasing up to 0.860 ± 0.003 for the case of top 5 crops being considered.

Chapter 6

Conclusion

We were initially faced with the problem of reviewing photos from a mobile phone photo collection, where the photos are used as wallpapers. This is a two-fold problem concerning the selection of photos to use as wallpapers, and centering or cropping the image appropriately such that important or salient areas are shown in an aesthetically pleasing way on the phone display.

When determining if an image could be used as a wallpaper, it is propagated through a deep convolutional neural network trained to identify object categories. The neuron activations together with annotations are then used to train a SVM. In practise this works quite well though with a bias towards distant natural scenes. Further work could be done in matching personal tastes.

When trained on the Michael dataset and evaluated on the Wookie dataset, an error rate of 3.7% is observed where most images are classified correctly. It is noted that for each annotator there is usually an improvement in precision when training on just his/her own annotations. This is to be expected as personal tastes would be better reflected in the learning phase. For the case of low correlation between annotations however, it would be interesting to combine annotations selectively to yield a more aesthetically pleasing final set of images.

Automatic cropping was selected for retargeting images to a specific aspect ratio. For cropping any given image, three cues are considered. These are: saliency composition, boundary simplicity, and content preservation. A SVM is trained with features based on these cues using a dataset sourced using the Reddit web service. The resulting model and algorithm works quite well, yielding a median maximum overlap of 0.782 for top 1 crops. This is an improvement over previous implementations.

Much more could be done to improve the suggested algorithms.

For instance, the datasets used in the selection stage could be improved with a greater variety of photos as well as a larger number of both photos and annotations. Annotators could be asked to annotate based on several keywords or themes, allowing for a classifier which attempts to adhere to user tastes. More features could be added during the learning stage. For instance, the colour distribution or blurriness of the image may change how suitable an annotator finds the image.

The cropping algorithm shows issues especially in the case where the saliency map or gradient map does not work as expected. Future work on enhancing algorithms for generating the two maps should improve the cropping algorithm. Further on, image segmentation using SLIC superpixels for example could allow for more advanced ways in respecting boundary simplicity as well as generating better crop boundaries in general.

Appendix A

Source Code

The following dependencies are required for running all code:

- `CMake 2.8+`
- `C++`
- `Boost 1.5+`
- `cvmatio - https://github.com/hbristow/cvmatio`
- `Python 2.7+`
- `scikit-learn`
- `numpy`
- `PIL`

`datasets/get_all_datasets.sh` retrieves all required data,
`scripts/trainer_pipeline.sh` trains the cropping model,
`scripts/get_features.py` calculates object class features for selection,
`scripts/train.py` trains the selection model,
and finally `scripts/pipeline.py` shows an example wallpaper slideshow where MMR is used to diversify displayed images.

All source code can be found at `https://github.com/swook/autocrop`.

Bibliography

[1] Chen Fang, Zhe Lin, Radomír Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the ACM International Conference on Multimedia*, pages 1105–1108. ACM, 2014.

[2] Jaesik Park, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Modeling photo composition and its application to photo re-arrangement. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2741–2744. IEEE, 2012.

[3] Jianzhou Yan, Shunjiang Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 971–978. IEEE, 2013.

[4] Pinaki Sinha, Hamed Pirsiavash, and Ramesh Jain. Personal photo album summarization. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 1131–1132. ACM, 2009.

[5] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *ACM Transactions on graphics (TOG)*, volume 26, page 10. ACM, 2007.

[6] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Multi-operator media retargeting. In *ACM Transactions on Graphics (TOG)*, volume 28, page 23. ACM, 2009.

[7] Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. In *ACM transactions on graphics (TOG)*, volume 29, page 160. ACM, 2010.

[8] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, pages 469–478. Wiley Online Library, 2010.

[9] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma. Auto cropping for digital photographs. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 4–pp. IEEE, 2005.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[12] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 153–160. IEEE, 2013.

[13] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM Press.